Demonstrating Lustre over a 100Gbps Wide Area Network of 3,500km

Robert Henschel*, Stephen Simms*, David Hancock*, Scott Michael*, Tom Johnson*, Nathan Heald*, Thomas William†, Donald Berry*, Matt Allen*, Richard Knepper*, Matthew Davy*, Matthew Link*, and Craig A. Stewart*
*Indiana University, 2709 E. Tenth Street, Bloomington, IN 47408
Tel.: +1-812-856-7511
Email: henschel@iu.edu
† TU Dresden - Center for Information Services and High Performance Computing
Zellescher Weg 12-14
01062 Dresden, Germany

***Abstract*—As part of the SCinet Research Sandbox at the Supercomputing 2011 conference, Indiana University (IU) demonstrated use of the Lustre high performance parallel file system over a dedicated 100 Gbps wide area network (WAN) spanning more than 3,500 km (2,175 mi). This demonstration functioned as a proof of concept and provided an opportunity to study Lustre's performance over a 100 Gbps WAN. To characterize the performance of the network and file system, low level iperf network tests, file system tests with the IOR benchmark, and a suite of real-world applications reading and writing to the file system were run over a latency of 50.5 ms. In this article we describe the configuration and constraints of the demonstration and outline key findings.**

I. Introduction

The SCinet Research Sandbox (SRS) at The International Conference for High Performance Computing Networking, Storage and Analysis 2011 (SC11) encouraged institutions to showcase new and innovative technologies in the area of networking. For the demonstration SCinet, in collaboration with ESnet and Internet2, provided all SRS participants with a 100 Gbps network connection from the SC11 show floor to the Internet2 backbone. Additionally, a 10 Gbps connection was provided to showcase software-defined networking via OpenFlow. The OpenFlow enabled link was used in a separate demonstration to show how the eXtensible Session Protocol can be used to create a control plane architecture for driving software defined networks [1].

The 100 Gbps link provided an end-to-end connection from the IU booth on the SC11 show floor to the IU Data Center in Indianapolis, Indiana. Figure 1 shows a network map of the link with the major routing points. Access to the network was governed by SCinet. Each participant was given time slots for exclusive use of the network. The slots were evenly distributed from Saturday, November 12th to Thursday, November 17th. In total, IU was provided nine test slots for a combined 16 hours. All testing that required access to the network links had to be performed during those times, from setting up the actual end-to-end network connectivity to performing file system and application tests. In addition, we were provided five demonstration slots for a total of four hours. Those time slots were used to showcase the capabilities of the system on the exhibition floor. All results described in this article were obtained in the 20 hours of test and demonstration time.

This paper is structured as follows: Section II outlines IU's work in the area of 100 Gbps networking and usage of the Lustre file system over wide area networks. Section III describes the hardware setup that was used for the SRS demonstration. Section IV outlines the network tests that we performed over the wide area link as well as local testing performed in our data center. This section also describes the network in more detail and outlines what was required to achieve good network performance. Section V describes the file system and application results. A summary and conclusion follows in section VI.

II. Related Work

In the Fall of 2005, IU received a National Science Foundataion Major Research Instrumentation award to fund construction and deployment of the Data Capacitor [2], a high bandwidth, large capacity facility for the short- to mid-term storage of scholarly data sets. Around that same time, four TeraGrid [3] sites began exploring using the Lustre file system over the wide area network. Indiana University, the National Center for Supercomputing Applications (NCSA), Oak Ridge National Laboratory (ORNL), and the Pittsburgh Supercomputing Center (PSC) began exporting, cross-mounting, and testing Lustre file systems between the sites. Using the Data Capacitor facility, performance testing between IU and PSC resulted in demonstrations of Lustre across the wide area network at the TeraGrid '06 conference. There it was shown that by properly tuning the TCP buffers a host could achieve 90% of maximum theoretical performance for reads and writes through a 1 Gbps Ethernet connection [4].

In 2006, one of the unique things about the Data Capacitor was the use of 10 Gbps interconnects for the object storage servers. As part of the Data Capacitor grant, 10 Gbps lines were run to three campus laboratories to provide fast transfer of instrument data to the Lustre file system. Following the 1 Gbps experiments with the TeraGrid network and our

SC12, November 10-16, 2012, Salt Lake
978-1-4673-0806-9/12/$31.00 ©2012 IEEE

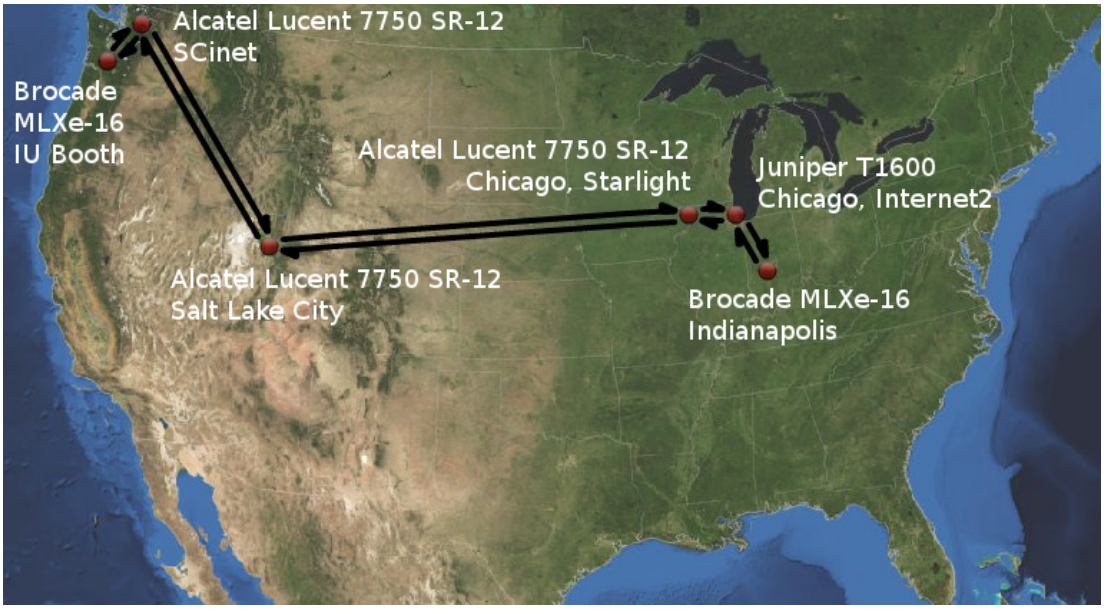

Fig. 1. Network map as seen using IU GlobalNOC World View. The end-to-end 100 Gpbs connection from the IU data center to the SC11 show floor comprises six hops in Indianapolis, Chicago, Salt Lake City, and Seattle.

successes using 10 Gbps links across campus, we decided to perform studies of 10 Gbps performance across longer distances. In early 2007 we showed that a single 10 Gbps client at ORNL could achieve a 977 MB/s transfer rate (out of a theoretical peak of 1250 MB/s) to the file system at IU [5], [6]. Consequently, users could access a familiar file system interface to mange their files and achieve rapid transfer of data across long distances without having to memorize complicated command line syntax or utilize an unreliable data transfer tool.

In the Fall of 2007, a team led by Indiana University demonstrated Lustre performance across a wide area network as part of the Bandwidth Challenge at SC07, in which competitors were asked to maximize the utility of a single 10 Gbps connection. Rather than demonstrate standard file system benchmarks, the team chose to showcase real world computation across distance. The demonstration supported five simultaneous scholarly workflows using the Data Capacitor, a smaller system constructed on the exhibit floor in Reno, Nevada, and resources from multiple institutions including Technische Universität Dresden which provided 4 Gbps over a 130 ms link from Germany. Running applications, the team achieved a peak bi-directional data transfer rate of 18.21 Gbps and a sustained data transfer rate of 16.2 Gbps, out of a possible 20 Gbps [7].

In order to make a wide area Lustre file system stretch beyond a demonstration and prove useful in a production environment, it was necessary to ensure that access to files and permissions could be managed properly. In 2007, Indiana University developed a set of patches to the Lustre code tree that would provide the mapping functionality needed to permit Lustre to bridge heterogeneous namespaces. [8]

The experience with Lustre over the WAN, and proven success at SC07 led to more extensive partnerships in WAN file system experimentation. Technische Universität Dresden partnered with IU in 2010 to perform file system tests across a 100 Gbps testbed that connected the cities of Dresden and Freiberg in Germany. This first ever commercial 100 Gbps link spanned the distance between TU Dresden and Bergakademie Freiberg with one Alcatel-Lucent switch at each site [9]. This distance is about 37 km geographically, but the fiber run from point to point was about 60 km. To simulate even longer distances, canisters of optical fiber cable were connected in series to create physical networks of 200 and then 400 km. Using the Lustre file system over the 60 km testbed, we achieved a unidirectional throughput of 11.8 GB/s and bidirectional throughput of 21.9 GB/s [10]. This is equivalent to network bandwidth of 94.4 Gbps unidirectional and 175.2 Gbps bidirectional. In other words, the 100 Gbps testbed link was effectively saturated.

In the United States, Internet2 has begun deploying a 100 Gbps Ethernet backbone at the direction of the Architecture and Operations Advisory Council. IU has been involved in this deployment because of its close ties to Internet2, functioning as the Network Operations Center, and having a seat on the board of directors. The project was funded through the second round of the Broadband Technologies Opportunities Program in early 2010. One of the goals of this project is the extension of Internet2's existing optical platform by upgrading the routing infrastructure to 100 Gbps. Phase one completed 100 Gbps optical deployment across the central area of the United States in early 2011, that included the first 100 Gbps Internet Protocol circuit between New York and Washington D.C. Phase two began in August 2011, connecting the eastern half of the U.S. through the Rocky Mountain regions to the west coast. Phase

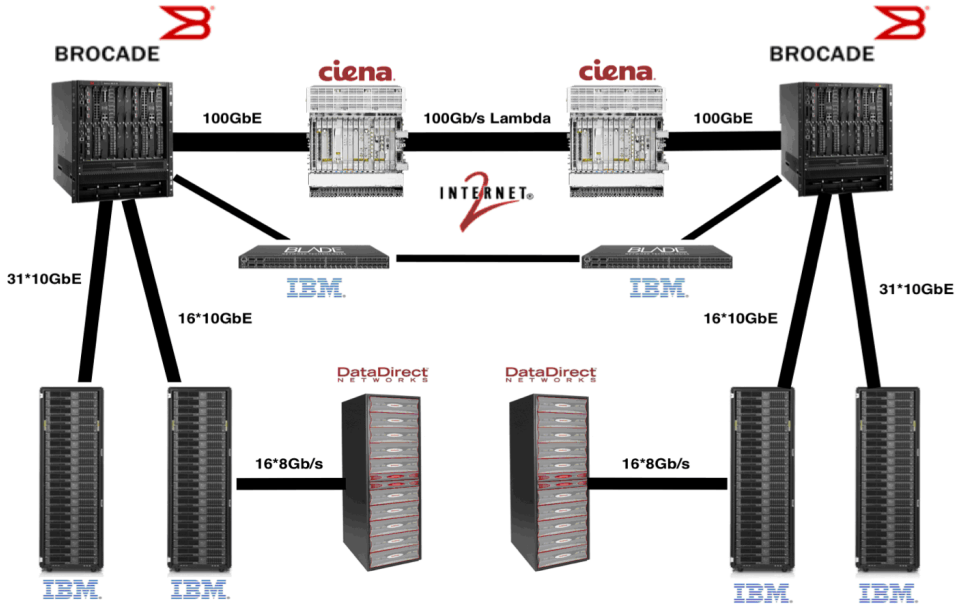

Fig. 2. Hardware setup

two has been completed, phase three will take Internet2 the remainder of 2012 and the first half of 2013 to complete. The phase one and two 100 Gbps deployments, in conjunction with ESnet's deployment of Advanced Networking Initiative, were critical for this demonstration. We utilized this path across the central U.S. from Chicago, IL to Seattle, WA.

III. Hardware Setup

Figure 2 shows the final configuration that was used between Seattle and Indianapolis. IBM provided 31 servers that functioned as compute nodes as well as the 16 storage servers that were attached to the DataDirect Networks (DDN) storage devices. Brocade and Ciena provided the routing equipment that enabled the network link from the show floor to Indianapolis.

The configurations of the compute cluster, storage, and networking components were identical in both Indianapolis and Seattle. The core networking component at each endpoint was a Brocade MLXe-16 router that provided a 100 Gbps Ethernet connection to the Ciena optical terminal managed by Internet2. The core router also provided a 10 Gbps link to an IBM BNT G8264 OpenFlow enabled switch at each endpoint. These switches were connected at 10 Gbps over a separate Internet2 connection. The 31 compute servers and 16 Lustre storage servers were attached directly to the Brocade core router at 10 Gbps using Twinax cables and Brocade 1860 dual-port adapters.

The compute servers were IBM System x iDataPlex dx360 M3 systems, each configured with dual Intel Xeon E5645 6-core 2.40 GHz processors, 24 GB of DDR3 RAM, a Brocade 1860 adapter, and a 250 GB SATA hard drive. The object storage servers (OSS) were IBM System x iDataPlex dx360 M3 servers each configured with an Intel Xeon E5645 6-core 2.40 GHz processor, 48 GB of DDR3 RAM, a Brocade 1860 adapter, and a 1 TB SATA hard drive. The OSS nodes at each site were connected directly to a DDN SFA10000 via 8 Gb Fibre Channel (FC). The SFA10000 drove five 60-slot storage enclosures configured with 2 TB SATA disk drives. The metadata server was identical to the compute servers, except it had 96 GB of RAM and was directly connected to a DDN EF3015 RAID system that contained twelve 300 GB 15K RPM SAS disk drives for Lustre metadata.

Due to space constraints on the show floor, server density was important. The dual-port Brocade 1860 adapters were able to saturate the 8 Gb FC links as well as the 10 Gbps Ethernet links simultaneously, allowing the use of rack dense IBM iDataPlex servers with only one PCI-Express slot available. The throughput of the DDN SFA10000 allowed us to use a single storage system for the Lustre OSS nodes at each site.

IV. Network and Network Tests

Before shipping equipment to the show floor in Seattle to participate in the SRS, the demonstration configuration was assembled in our data center for testing. There we performed tests with the two Brocade MLXe routers connected back to back with the 100 Gbps connection spanning a few meters. Local network tests across this connection showed a round trip time of 0.24 ms using an ICMP echo request (ping) and maximum stable performance of 98 Gbps using TCP iperf [11] and 9.6 GB/s using IOR [12].

The network link that was used for this demonstration provided end-to-end 100 Gbps connectivity from the IU booth on the SC11 show floor in Seattle to the data center in Indianapolis. In Seattle and on the show floor, SCinet was responsible for the network link. The segment from Seattle to Salt Lake and on to Chicago is now being integrated into

TABLE I
KEY RESULTS OF THE DEMONSTRATION. PERCENTAGE OF THEORETICAL PEAK IS GIVEN IN PARENTHESIS.

	Compute nodes to local storage over 100G	**From Seattle compute nodes to Indianapolis storage**
Round trip time (ping)	0.24 ms	50.5 ms
TCP iperf, one stream	9.8 Gbps (98%)	9.8 Gbps (98%)
TCP iperf, 30 streams (peak)	98 Gbps (98%)	96 Gbps (96%)
IOR, 1 client, 8 servers	1.2 GB/s (96%)	1.2 GB/s (96%)
IOR, 30 clients, 16 servers	9.6 GB/s (77%)	6.5 GB/s (52%)
8 Applications concurrently (peak)	8.8 GB/s (70%)	6.2 GB/s (50%)
8 Applications concurrently (sustained)	7.9 GB/s (63%)	5.6 GB/s (45%)

the Internet2 network, we were using a very early build out of the soon to be production network. It should be noted that especially the connection from Salt Lake City to Seattle had been established at 100 Gbps only two weeks prior to SC11. For the segment from Chicago to Indianapolis, we worked with Internet2 to upgrade our existing connection to 100 Gbps. As of March 2012, this link is in production.

Once the link was established, we measured a round trip time of 50.5 ms between the compute cluster in Seattle and the storage cluster in Indianapolis using ping. Initial TCP iperf tests had shown good performance for two parallel streams, each at 10 Gbps. However, when adding more iperf streams to the link, performance dropped and throughput become unstable. We tuned two specific parts of the network to arrive at a stable 80 Gbps throughput with TCP iperf. Fist we removed a link aggregation group on both sides and second we statically mapped two VLANs onto the two 50 Gbps connections that made up the 100 Gbps link. Both steps are discussed in more detail in the following paragraph.

Our initial hardware setup included a Brocade VDX switch between the compute and storage nodes and the Brocade MLXe router. The VDX was used to aggregate the 47 server links into 12 that were then routed by the MLXe into the Ciena optical terminal. Despite our best efforts and assistance from Brocade engineers we were unable to completely eliminate the performance impact of the VDX's link aggregation. Directly connecting all compute and storage resources into the MLXe router eliminated the layer of link aggregation that had been present on both sides of the connection [13]. In addition, we worked with ESnet to optimize the configuration of the three Alcatel-Lucent 7750 SR-12 routers in the path from Seattle to Chicago. Those routers partition the 100 Gbps network traffic onto two 50 Gbps connections. Since our end-to-end connection was mapped using two VLANs, we changed the configuration of those switches to statically map one VLAN per 50 Gbps connection. Each VLAN contained half of the cluster nodes. Similar tuning was performed for the connection from Chicago to Indianapolis, involving a Juniper T1600 and a Brocade MLXe router. Following all of those modifications, we were able to put a stable 80 Gbps of TCP iperf traffic on the link, 40 Gbps per 50 Gbps connection and VLAN. This was achieved using eight servers on both sides of the link, four per VLAN. Adding another single 10 Gbps stream to either of the connections resulted in congestion and degraded performance. We then discovered that the two 50 Gbps connections were rated at slightly less than 50 Gbps, which explained the congestion when adding the fifth server. However, when oversubscribing the link by using all 30 compute nodes on each end, we were able to achieve a peak throughput of 96 Gbps. Those tuning steps yielded a stable connection only in one direction, from Seattle to Indianapolis. We applied the same tuning to the return path, but the performance did not improve in a similar manner. We continued to see massive packet reordering. Since 11 of our available 16 test hours had already been used, we decided to focus our efforts on data transfers in just one direction. It is hard to quantify the performance impact of the packet reordering of the return path. Ideally, only very small requests are sent in this direction, starting or acknowledging requests. Packet reordering should affect such traffic less than bulk data traffic.

Listing 1. Tuning parameters for the network

```
net.ipv4.tcp_rmem=4096 65536 167772160
net.ipv4.tcp_wmem=4096 65536 167772160
net.core.rmem_max=167772160
net.core.wmem_max=167772160
net.core.netdev_max_backlog=30000
eth2 txqueuelen 10000
eth2 mtu 9000
FlowControl off
```

Listing 1 shows the tuning parameters that were set on all nodes. We increased the maximum TCP buffer to 167 MB and increased the sending and receiving queues to 10000 and 30000. In addition, we enabled MTU 9000 and configured all nodes to use the `tcp_bic` network stack. Flow control was disabled on the network adapters and the routers.

V. LUSTRE AND APPLICATIONS

Once the TCP iperf tests showed good performance in one direction, we proceeded to export the Lustre file system over the WAN by mounting the Indianapolis-based file system on the compute nodes in Seattle. We used the IOR benchmark to characterize the performance of the file system. Due to time constraints, we primarily tested write operations with IOR.

Table I shows the summary of our performance tests, listing the peak TCP iperf results for one and 30 streams, the peak IOR results for one client and for all 30 clients, and the average and peak throughput of running eight applications concurrently distributed to 30 clients. The table compares the results that we measured over the WAN with measurements that were taken

TABLE II
SINGLE NODE IOR TESTS, USING 8 IOR PROCESSES.

Lustre RPC	Lustre Stripe Count	Local Throughput MB/s	WAN Throughput MB/s
32	32	1238	1159
32	16	1240	1174
32	8	1221	1176
16	32	1221	1124
16	16	1192	1137
16	8	1196	1110
8	32	1010	683
8	16	1063	860
8	8	1058	830

prior to shipping the equipment to Seattle, when the whole system was assembled in the IU machine room. The setup in both cases was identical, except for the local test case where the two Brocade MLXe routers were directly connected to each other.

We first tested the peak performance of a single node. Table II shows detailed results from IOR tests that were performed both local and over the WAN. In the local test case, we achieved 1.2 GB/s throughput running eight IOR processes on a single node, each of them writing its own file using 1 MB blocks. In aggregate 54 GB of data were written, more than twice the available memory in the compute nodes. We did not use the IOR option `O_direct`, as we wanted to compare IOR performance to subsequent application performance and none of the tested applications supported direct I/O. On the file system level, we set the Lustre file striping to 8, 16, and 32, meaning the data for each file is distributed evenly across 8, 16, or 32 OSSs. The Lustre parameter RPCs in flight [14] was also set to 8, 16, or 32. This represents the number of outstanding write requests per OSS. In the local case, we achieved good performance with either 16 or 32 RPCs in flight. When repeating the tests over the WAN we also achieved a peak of 1.2 GB/s. In the WAN environment it is crucial to increase the number of RPCs in flight from 16 to 32. Higher numbers of RPCs in flight will account for the higher latency of the connection between the compute nodes and the storage system. Comparing the local measurements with the measurements over WAN we found almost no difference for IOR on a single node.

We then used IOR to measure the I/O performance of all 30 compute nodes together. In the local test case, we achieved 9.6 GB/s when using three IOR processes per compute node, each of them writing its own file. For Lustre we used a striping of 8 and an RPC in flight setting of 8. When performing this test over the distance, we increased the RPC in flight setting to 32, while leaving the rest of the configuration unchanged, and measured 6.5 GB/s. The local measurements were more than 3 GB/s faster than the measurements over the WAN. We believe that with more Lustre tuning the results could be improved. The only tuning that we were able to apply in the short time available was increasing the RPCs in flight setting. There are other Lustre tuning parameters that can be used to further improve performance over the WAN. For example, adjusting the `max_dirty_mb` [14] setting on the clients would allow for more client side caching and thus increase the blocks of data that are sent back to the server at a given point in time. Increasing the `peer_credits` [15], [16] parameter would increase the threads available to the lustre networking subsystem and could support higher values for RPCs in flight. It should also be noted, as mentioned in section 4, the return connection from Indianapolis to Seattle was not configured optimally during our tests. This impacted reads from the Indianapolis file system and acknowledgement packets for writes.

The next step was running real world applications on the Seattle compute nodes that mounted the Lustre file system from Indianapolis. Table III shows the eight I/O intensive applications that were run concurrently. Those applications are normally either running on the HPC systems at IU or were contributed by collaborators.

Heat3D is a code that is used to simulate heat diffusion. The code was developed at IU and for this demonstration we configured it with a model of a CPU heat sink with 5 fins. At the beginning, the code reads the input configuration file as unformatted FORTRAN I/O and then writes out checkpoint files. VampirTrace is part of the Vampir toolchain that is developed at the Center for Information Services and High Performance Computing (ZIH) at Technische Universität Dresden, Germany. The tool was used to trace the runtime behavior of the Heat3D code. It was configured to record OpenMP events from the Heat3D code, and write them to one trace file per OpenMP thread. Enzo is an adaptive mesh refinement, grid-based code, which is designed to do simulations of cosmological structure formation. It can also be used for astrophysical fluid dynamics simulations more generally, and has been used to simulate star formation and galactic clusters. Enzo reads an initial configuration file and then writes out checkpoint files. The National Center for Genome Analysis Support (NCGAS) contributed a workflow that aligns a dataset to a reference genome and then calculates the single nucleotide polymorphisms. Novoindex and Bowtie were used in the workflow. All tools in this workflow perform text I/O in very small chunks. The Ocean Land Atmosphere Model (OLAM) was used with a global climate benchmark. This work is part of a collaboration with the University of Miami through the XSEDE Extended Collaborative Support

TABLE III
APPLICATIONS

Application/Workflow	Domain	Number of Nodes
Heat3D	Heat Diffusion	8
VampirTrace [17]	Application Performance Analysis	7
Enzo [18]	Astronomy	6
NCGAS [19]	Genomics	3
OLAM [20]	Weather	2
CMES [21]	Computational Neuroscience	2
Gromacs	Molecular Dynamics	1
ODI-PPA [22]	Astronomy	1

Service program. OLAM also reads an initial configuration and then performs checkpointing. The Computational Model for Electroencephalography responses in Schizophrenia (CMES) simulates a network of neurons to explain the alteration in EEG responses in schizophrenia. Gromacs was used to study the structural change of the CsoR copper sensing protein when it binds to copper. The work is part of an ongoing collaboration with the chemistry department at IU. After reading in the start configuration, Gromacs writes out intermediate configurations in large blocks. The One Degree Imager Pipeline Portal and Archive (ODI-PPA) project contributed a Tier 1 workflow that will be used to reduce images produced by the ODI instrument. ODI-PPA is a developed out of a collaboration with the WIYN Consortium, comprised of the University of Wisconsin, Indiana University, Yale University and the National Optical Astronomy Observatory (WIYN). Due to the limited execution time, only the first step of the workflow was run, consisting of copying files from the simulated remote location to the local cluster. VampirTrace, NCGAS, Gromacs and ODI-PPA are already using the Data Capacitor, available on all IU HPC resources and mounted at various other HPC centers in the US. We configured each of the applications to use a subset of the 30 compute nodes; the number of nodes per application is shown in the last column of the table. Running the applications locally yielded an average throughput of 7.9 GB/s. When running the applications over the distance we saw an average throughput of 5.6 GB/s. In both cases, the peak throughput was more than 0.5 GB/s above the average throughput, due to the cyclical I/O behaviour of the applications. We employed the same Lustre tuning that was used to run IOR. The performance difference between the local execution and the execution over the WAN is more than 2 GB/s. The same factors that influenced the performance of IOR also influenced the applications. Since most of the applications also needed to read data from the file system the poor performance of the Indianapolis to Seattle link played a larger role for applications.

Table I also shows the percentage of peak for the results, both in the local and the WAN test case. For TCP iperf tests, we see essentially no difference between local and WAN. For IOR, we see almost no difference for the single node test. However, when running on all 30 clients, the local measurements are a lot better, 77% vs. 52%. For applications, the same holds true. The local sustained throughput is at 63% of peak while the WAN one is only at 45%. The TCP iperf and the IOR values can be compared to results from the 60 km 100 Gbps test bed at ZIH in Germany [10]. In this test bed, a very similar configuration was used, with the exception that more compute nodes were available, a lot more time was available to properly configure and tune all parts of the link and the test bed was only 60 km. The round trip time at ZIH was 0.72 ms. For TCP iperf, 98.78 Gbps (98.8% of peak) was achieved and for IOR on Lustre, 11.8 GB/s (94.4% of peak) was measured. Those values certainly present a realistic upper bound for what can be achieved over a 100 Gbps WAN link.

VI. CONCLUSION

This demonstration continues IU's long history of bridging the Lustre file system from HPC systems in the data center to systems on campus and further out to resources at national centers and other universities. Compared to similar demonstrations in the past [7], this one used network technology that had only recently been introduced, and deployed in a production environment over such long distances. This is especially reflected in the amount of low level network tuning that was required before we could deploy the Lustre file system and the fact that, in the end, only unidirectional traffic worked properly.

By performing this demonstration and utilizing the soon to be production ESnet/Internet2 infrastructure we learned two key things about how Lustre performs over a 100 Gbps WAN. The first, in fact more of a lesson of 100 Gbps networks in general, is that link aggregation and VLANs can have a significant impact on performance if clients are not configured and placed correctly. Much of our tuning and troubleshooting efforts were devoted to understanding how network traffic was being routed over the 100 Gbps link. This was then followed by devising the proper mapping of clients to network paths to produce traffic that would maximize throughput while not causing contention issues. Our previous experience with 10 Gbps networks did not require shaping of traffic in order to achieve good performance. Unfortunately, it seems, for now at least, that fully utilizing cross-country 100 Gbps networks is not as simple as plugging many 10 Gbps clients into a 100 Gbps endpoint.

The other, and perhaps more surprising finding, is how dramatically complications with link aggregation in the network can affect Lustre-WAN performance. It is clear that Lustre can deliver excellent performance over 100 Gbps networks, in both LAN and WAN settings (see section II). However, both the local tests performed in our data center and at ZIH

took place in a much more controlled networking environment, with direct switching and on shared subnets. One might argue that the decrease in performance is due to increased network latency, but Lustre performance remained roughly consistent over an, albeit small, range of latencies. The key discriminator was not latency, but complexity. The relatively simple, well-defined and controlled networks in our data center and the ZIH configurations became a complex series of routing and link aggregation for the SRS demonstration. This change resulted in efficiencies dropping from well over 90% to 50%. Of course, the studies described in section II along with the SRS demonstration used a relatively small number of clients attempting to saturate a 100 Gbps link. Perhaps a system or systems utilizing more 10 Gbps clients could achieve higher efficiencies.

Experimentation enabled by the SRS pushed the limits of the first 100 Gbps WAN deployments in the short time allotted, without such a venue to conduct experiments the transition from pre-production to high performing production research and education networks would be much more difficult. Our SRS entry was as much an experiment at the network layer as it was at the file system or applications layer. This is substantiated by the fact that across the SRS we observed 100 Gbps equipment from three different vendors interoperating for the first time across thousands of miles. Acceptable bidirectional performance had not been achieved by the end of the conference, even with experts from multiple vendors and institutions working on the link. As both 100 Gbps technology and national deployments mature, issues like those we encountered at the SRS demonstration will be addressed and overcome. Vendors are already working on the second generation of 100 Gbps equipment. The new equipment will feature a unified 100 Gbps fabric, eliminating the need to split the signal between two 50 Gbps channels and thus requiring statically mapped VLANs. This will greatly simplify managing high bandwidth connections in the national backbone. A second important trend are lower cost optical elements and broader support for standard form factors among all vendors. This will allow for better interoperability between vendors.

It should be noted that even though we were unable to fully saturate the SRS 100 Gbps link, real world applications were still able to run over the WAN and, on average, each achieve several Gbps throughput. Indeed, from the end user's point of view, the true breakthrough of Lustre-WAN is not necessarily the ability to achieve 90%+ network efficiency, though it has been shown to be possible, but the streamlining of their workflow by eliminating the need for user-directed file transfers [23], [24]. Overall, we see the future of Lustre-WAN and 100 Gbps networking to be very promising, albeit with a period of adjustment in the near term future.

ACKNOWLEDGEMENTS

The authors would like to thank the contributions of our collaborators and technical staff. Real world application tests were provided by the following individuals: CMES - Chathura Herath and Suresh Marru (Indiana University), Enzo - Britton Smith (University of Colorado, Boulder) and Matthew Turk (Columbia University), Gromacs - Abhigna Polavarapu (Indiana University), NCGAS - Le-Shin Wu (Indiana University), ODI-PPA - Arvind Gopu (Indiana University), OLAM - Yu Ma (Indiana University). The following staff at IU provided excellent support in the administration of the systems, testing applications, visualization, technical writing, and project management: Edward Balas, Janae Cummings, Kurt Seiffert, Daphne Siefert-Herron, Bill Sherman, Martin Swany, Jenett Tillotson, George Turner, and Josh Walgenbach. Internet2 provided network setup and support: Andrew Lee, Chris Robb, and Matthew Zekauskas. ESnet staff also provided setup and troubleshooting assistance: Evangelos Chaniotakis and Patrick Dorn.

The following vendors provided loaner hardware to support this project: Brocade loaned all the server host adapters, two MLXe-16 100 Gbps switches, required Fibre Channel and Ethernet optics, and all the Twinax cables used to connect the servers to the network. Ciena loaned optical equipment to Internet2 to enable the 100 Gbps network from Chicago to Seattle. DataDirect Networks loaned two SFA10000 storage systems for Lustre object storage targets and two EF3015 RAID arrays for Lustre metadata storage. IBM contributed 22 iDataPlex dx360 M3 servers for the project as well as two OpenFlow enabled BNT G8264 switches with required optics. Internet2 provided networking equipment to extend their 100 Gbps network from Chicago to Indianapolis.

REFERENCES

[1] E. Kissel, G. Fernandes, M. Jaffee, M. Swany, and N. Zhang, "Driving Software Defined Networks with XSP," in *In submission for SDN'12: Workshop on Software Defined Networks*. IEEE, 2012.

[2] C. Stewart, R. Bramley, C. Pilachowski, B. Plale, and T. J. Hacker, "MRI: Acquisition of a High-Speed, High Capacity Storage System to Support Scientific Computing: The Data Capacitor." [Online]. Available: http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0521433

[3] C. Catlett, W. E. Allcock, P. Andrews, R. Aydt, R. Bair, N. Balac, B. Banister, T. Barker, M. Bartelt, P. Beckman, F. Berman, G. Bertoline, A. Blatecky, J. Boisseau, J. Bottum, S. Brunett, J. Bunn, M. Butler, D. Carver, J. Cobb, T. Cockerill, P. F. Couvares, M. Dahan, D. Diehl, T. Dunning, I. Foster, K. Gaither, D. Gannon, S. Goasguen, M. Grobe, D. Hart, M. Heinzel, C. Hempel, W. Huntoon, J. Insley, C. Jordan, I. Judson, A. Kamrath, N. Karonis, C. Kesselman, P. Kovatch, L. Lane, S. Lathrop, M. Levine, D. Lifka, L. Liming, M. Livny, R. Loft, D. Marcusiu, J. Marsteller, S. Martin, S. McCaulay, J. McGee, L. McGinnis, M. McRobbie, P. Messina, R. Moore, R. Moore, J. Navarro, J. Nichols, M. E. Papka, R. Pennington, G. Pike, J. Pool, R. Reddy, D. Reed, T. Rimovsky, E. Roberts, R. Roskies, S. Sanielevici, J. R. Scott, A. Shankar, M. Sheddon, M. Showerman, D. Simmel, A. Singer, D. Skow, S. Smallen, W. Smith, C. Song, R. Stevens, C. Stewart, R. B. Stock, N. Stone, J. Towns, T. Urban, M. Vildibill, E. Walker, V. Welch, N. Wilkins-Diehr, R. Williams, L. Winkler, L. Zhao, and A. Zimmerman, "TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications," in *High Performance Computing and Grids in Action*, L. Grandinetti, Ed. Amsterdam: IOS Press, 2008, pp. 225–249.

[4] S. Simms, B. Hammond, M. Link, and C. Stewart, "The Data Capacitor Project," demonstration given at TeraGrid 06 conference, 2006.

[5] S. C. Simms, G. G. Pike, and D. Balog, "Wide area filesystem performance using lustre on the teragrid," in *TeraGrid'07 Conference*, 2007.

[6] S. C. Simms, G. G. Pike, S. Teige, B. Hammond, Y. Ma, L. L. Simms, C. Westneat, and D. A. Balog, "Empowering distributed workflow with the data capacitor: maximizing lustre performance across the wide area network," in *SOCP '07: Proceedings of the 2007 workshop on Service-oriented computing performance: aspects, issues, and approaches*. New York, NY, USA: ACM, 2007, pp. 53–58.

[7] "Indiana University. Team Led by IU Wins Supercomputing Bandwidth Competition." [Online]. Available: http://newsinfo.iu.edu/news/page/normal/6839.html

[8] J. Walgenbach, S. C. Simms, K. Westneat, and J. P. Miller, "Enabling Lustre WAN for production use on the TeraGrid: a lightweight UID mapping scheme," in *Proceedings of the 2010 TeraGrid Conference*, ser. TG '10. New York, NY, USA: ACM, 2010, pp. 19:1–19:6.

[9] HPCWire, "Indiana University Announces Saturation of First Commercial 100 Gigabit Link," 2010. [Online]. Available: http://www.hpcwire.com/hpcwire/2010-11-15/indiana_university_announces_saturation_of_first_commercial_100_gigabit_link.html

[10] M. Kluge, S. Simms, T. William, R. Henschel, A. Georgi, C. Meyer, M. S. Mueller, C. A. Stewart, W. Wnsch, and W. E. Nagel, "Performance and quality of service of data and video movement over a 100 gbps testbed," *Future Generation Computer Systems*, no. 0, pp. –, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X12001380?v=s5

[11] iperf Team, "Home page," http://sourceforge.net/projects/iperf/, 2012.

[12] H. Shan and J. Shalf, "Using IOR to analyze the I/O performance for HPC platforms," Lawrence Berkeley National Laboratory, Tech. Rep., 2007, http://escholarship.org/uc/item/9111c60j.

[13] R. Knepper, S. Michael, W. Johnson, R. Henschel, and M. Link, "The lustre file system and 100 gigabit wide area networking: An example case from sc11," in *Networking, Architecture and Storage (NAS), 2012 7th IEEE International Conference on*, June 2012, p. in press.

[14] "Lustre 1.8 Operations Manual." [Online]. Available: http://wiki.lustre.org/manual/LustreManual18_HTML/LustreProc.html#50651263_pgfId-1290515

[15] "Lustre Development - Credits, peer credits and concurrent sends." [Online]. Available: http://comments.gmane.org/gmane.comp.file-systems.lustre.devel/2041

[16] S. Michael, L. Zhen, R. Henschel, S. Simms, E. Barton, and M. Link, "A study of lustre networking over a 100 gigabit wide area network with 50 milliseconds of latency," in *Proceedings of the fifth international workshop on Data-Intensive Distributed Computing Date*, ser. DIDC '12. New York, NY, USA: ACM, 2012, pp. 43–52.

[17] H. Brunst, D. Hackenberg, G. Juckeland, and H. Rohling, *Comprehensive Performance Tracking with Vampir 7*, ser. Tools for High Performance Computing 2009. Springer Berlin Heidelberg, 2010, pp. 17–29, http://dx.doi.org/10.1007/978-3-642-11261-4_2.

[18] B. W. O'Shea, G. Bryan, J. Bordner, M. L. Norman, T. Abel, R. Harkness, and A. Kritsuk, "Introducing Enzo, an AMR Cosmology Application," *ArXiv Astrophysics e-prints*, Mar. 2004.

[19] "National Center for Genome Analysis Support." [Online]. Available: http://ncgas.org

[20] D. Medvigy, R. L. Walko, M. J. Otte, and R. Avissar, "The Ocean-LandAtmosphere Model: Optimization and Evaluation of Simulated Radiative Fluxes and Precipitation." *Mon. Wea. Rev.*, vol. 138, p. 19231939, 2010.

[21] G. P. Krishnan and M. Bazhenov, "Ionic dynamics mediate spontaneous termination of seizures and postictal depression state." *J Neurosci*, vol. 31, no. 24, pp. 8870–82, 2011.

[22] "The WIYN One Degree Imager." [Online]. Available: http://www.noao.edu/wiyn/ODI/

[23] R. Henschel, S. Michael, and S. Simms, "A distributed workflow for an astrophysical OpenMP application: using the data capacitor over WAN to enhance productivity," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, ser. HPDC '10. New York, NY, USA: ACM, 2010, pp. 644–650.

[24] S. Michael, S. Simms, W. B. Breckenridge, III, R. Smith, and M. Link, "A compelling case for a centralized filesystem on the TeraGrid: enhancing an astrophysical workflow with the data capacitor WAN as a test case," in *Proceedings of the 2010 TeraGrid Conference*, ser. TG '10. New York, NY, USA: ACM, 2010, pp. 13:1–13:7.